

Sequence signatures of genes with accompanying antisense transcripts in *Saccharomyces cerevisiae*

LI YingXue¹, LIU XueNing¹, WANG XiaoWo¹ & ZHANG XueGong^{1,2*}

¹MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China;

²School of Life Sciences, Tsinghua University, Beijing 100084, China

Received October 14, 2013; accepted November 22, 2013; published online December 23, 2013

Recent studies have found many antisense non-coding transcripts at the opposite strand of some protein-coding genes. In yeast, it was reported that such antisense transcripts play regulatory roles for their partner genes by forming a feedback loop with the protein-coding genes. Since not all coding genes have accompanying antisense transcripts, it would be interesting to know whether there are sequence signatures in a coding gene that are decisive or associated with the existence of such antisense partners. We collected all the annotated antisense transcripts in the yeast *Saccharomyces cerevisiae*, analyzed sequence motifs around the genes with antisense partners, and classified genes with and without accompanying antisense transcripts by using machine learning methods. Some weak but statistically significant sequence features are detected, which indicates that there are sequence signatures around the protein-coding genes that may be decisive or indicative for the existence of accompanying antisense transcripts.

antisense transcript, pattern recognition, motif finding, yeast

Citation: Li YX, Liu XN, Wang XW, Zhang XG. Sequence signatures of genes with accompanying antisense transcripts in *Saccharomyces cerevisiae*. Sci China Life Sci, 2014, 57: 52–58, doi: 10.1007/s11427-013-4597-7

Genome-wide existence of natural antisense transcription has been discovered in various species from yeast to human [1–8]. Among the 6575 annotated coding genes in the yeast genome, 531 genes have been reported to have accompanying antisense transcripts while the other 6044 do not [9]. For the convenience of the following description, we call them genes with and without antisense partners, respectively, and regard the corresponding protein-coding gene as the primary gene.

A recent study from Lars Steinmetz's lab reported that the antisense transcript works as an on-off switch on the regulation of the primary gene [10]. The on-off switch means that the regulatory effect can be turned on or off depending on the circumstances. The switch is turned on when

the coding gene is confronted of genetic and environmental changes and off under normal circumstances. The switch function also has the following characteristics: (i) The switch is threshold dependent. An antisense transcript usually inhibits its partner gene expression in low levels, and the primary gene in high expression level usually represses the antisense expression in return by forming a feedback loop. (ii) Presence of Transcription Start Site (TSS) overlapping makes the switch function more effective. "TSS overlapping" means that the transcription region of the primary gene covers the TSS of the antisense partner, or the transcription region of the antisense transcript covers the TSS of the primary gene. (iii) The interaction between the sense-antisense pair makes the primary gene have a larger expression changes in response to genetic and environmental changes.

*Corresponding author (email: zhangxg@tsinghua.edu.cn)

Previous studies mainly focused on the function of the antisense transcripts themselves. Since the major role of these partner antisense transcripts is to regulate their primary genes, it would be interesting to ask what determines whether a gene has such a partner or not? To the best of our knowledge, this question has not been documented in the literature. Therefore, we designed this study to investigate whether there are signals or at least signatures in the gene sequence that are associated with the presence or absence of antisense partners.

We collected all reported antisense transcripts in *Saccharomyces cerevisiae* and the sequences of genes with and without antisense partners. We first applied a pattern recognition method to study whether the sequence features of these two classes are separable. The classification accuracy was not high but was statistically significant according to permutation test. This indicates that the two groups are separable to some degree by sequence signatures. Then we did motif finding analysis to detect the possible sequence motifs which may be associated with the existence of the antisense transcripts. Some weak but significant sequence motifs like 5'-TGTTG-3' were found to be enriched in genes with antisense partners.

1 Data

1.1 Yeast genome annotation dataset

We downloaded the yeast genome annotation data (Version R64-1-1 released on April 13, 2013) in gff format from the SGD database (<http://www.yeastgenome.org/download-data/curation>). The data includes genomic sequences, coordinates and annotations.

1.2 Antisense transcript dataset

The antisense transcript dataset was obtained from [9], downloaded from the site <http://genomebiology.com/content/supplementary/gb-2010-11-3-r24-s3.xls>. This dataset contains 523 antisense transcripts detected in *S. cerevisiae* S288C strain. These transcripts were found to be expressed during the mitotic cell cycle and associated with 531 primary protein-coding genes (each antisense transcript has one or two corresponding primary genes). The data file provides us the chromosome ID, coordinate, length and the primary gene name of each antisense transcript. Although there are several other yeast antisense transcript datasets released by the same lab from experiments with different growing media or from different strains, the dataset we chose contains the largest collection and covers the majority of the antisense transcripts reported in the other datasets.

In this paper, we focused on the 514 antisense transcripts that are associated with protein-coding genes, and the other nine antisense transcripts accompanying micron genes or mitochondrial genes were not considered. As shown in Figure 1,

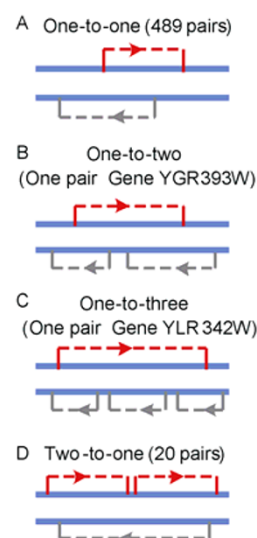


Figure 1 Four types of primary-antisense pairs. The primary gene is indicated by the red arrow, and its antisense partner is indicated by the gray arrow. A, One-to-one correspondence. Most pairs (489) are of this type. B, One-to-two: the gene YDR393W is accompanied with two antisense transcripts. We still use “pair” to refer to this partner group, as well as the groups in the situation of (C) and (D). C, One-to-three: the gene YLR342W is accompanied with three antisense transcripts. D, Two-to-one: two adjacent primary genes share one antisense partner transcript. Twenty antisense transcripts belong to this type.

there are four types of relative relations of the primary genes and their antisense partners according to the different number of involved genes or antisense transcripts.

1.3 Pre-processing of the original dataset

We got the mRNA annotation of all yeast genes from the database, and separated them into two groups: (i) genes with antisense partners and (ii) genes without antisense partners. Chromosome IDs and coordinates of TSSs and TESs (transcription end sites) were extracted from the annotation file, as well as the TSSs and TESs of the antisense transcripts. Genomic sequences were extracted according to the genomic coordinates of these sites.

To find the sequence signatures that may associate with antisense transcription, we extracted six types of sequences around primary genes and their antisense partners. Their positional relations are illustrated in Figure 2A, and the sequence-extraction regions of the gene YDL210W and its partner antisense transcripts are shown in Figure 2B as an example.

To study the possible transcriptional influence between the sense and antisense transcripts, we further subdivided the sense-antisense pairs according to the presence of TSS overlapping. We divided the 531 genes with antisense partners into two groups according to whether the TSS of the primary gene overlaps with the partner antisense transcript. We also divided the 514 antisense transcripts into two

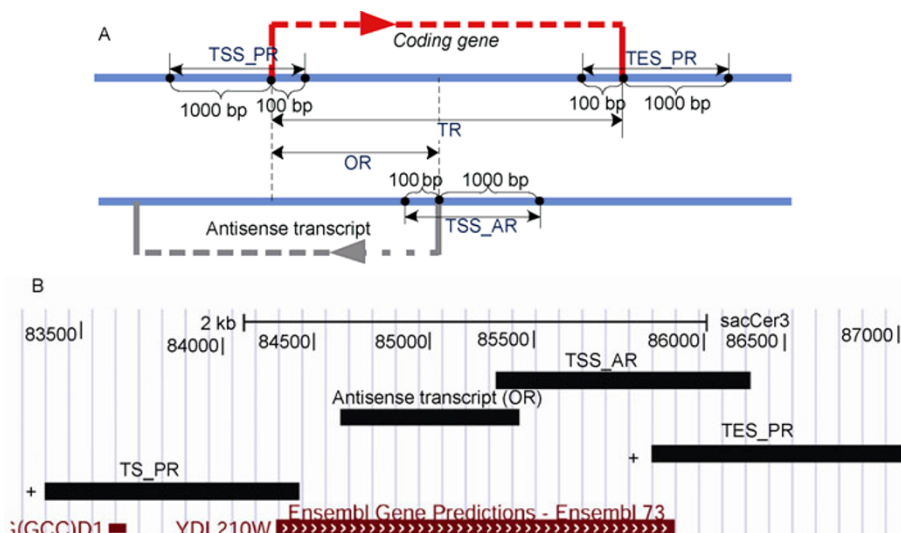


Figure 2 Regions of the sequences extracted from the primary coding gene and its antisense partner transcript. A, The general definition. B, The example of gene YDL210W. Notations: TR (transcription region), region of a gene from TSS to TES; OR (overlap region): overlapped region of the primary-antisense pair; TSS_PR: from upstream 1000 bp to downstream 100 bp of the TSS of the primary gene; TSS_AR: from upstream 1000 bp to downstream 100 bps of TSS of the antisense transcript; TES_PR: from upstream 100 bp to downstream 1000 bp of TES of the primary gene. In (B), the primary gene covers the whole region of the partner antisense transcript, therefore the whole antisense transcript region is the OR.

groups by checking whether the TSS of the antisense transcript overlaps with the corresponding transcript of the primary gene.

2 Methods

2.1 Pattern recognition

2.1.1 Features for the classification

We first designed the pattern recognition task of classifying the two classes of samples: genes with and without partner antisense transcripts. The two classes are also referred to as the foreground and background group in the motif finding work. The TR sequences of the primary genes were extracted as the original samples. We counted the occurrences of all k -mer nucleotide words (short words of k nucleotides) in each sample. We tried k from 2 to 5. The counts of the k -mer words composed the feature vectors for the classification, with dimension of 16, 64, 256 and 1024 corresponding to $k=2, 3, 4$ and 5, respectively. The feature vectors were normalized so that each element is the relative percentage of the occurrence of a k -mer word in the sample.

2.1.2 Fisher's linear discriminant analysis

We used the Fisher's linear discriminant analysis (FLDA) for the classification of genes with and without antisense partner transcripts. FLDA is a basic pattern recognition method for linear classification [11]. It projects the feature vectors of all samples to a linear direction along which the two classes are best separated while each of the classes is best condensed. Due to its simplicity and robustness, it is widely used as the first attempt on a new classification

problem. FLDA is chosen in this study because it is suitable for answering whether the two classes of genes (with and without antisense partner) can be separated with sequence signatures.

2.1.3 Experiment Settings

The two classes have 531 samples with antisense partner and 6044 samples without antisense partner, respectively. We did 20 rounds of pattern recognition experiments. In each round, 500 samples from both the foreground and background group were randomly selected to balance the two classes in sample size. We used 5-fold cross-validation to evaluate the result of the pattern recognition experiment. That is, we divided the 500+500 samples into five subsets of equal size (100+100 samples in each subset), used four subsets to train the FLDA method to obtain a classifier and used the remaining subset to test the trained classifier. Each subset was used as the test set once, and the test errors were averaged as the cross-validation error.

Permutation experiments were adopted to assess the significance of achieved error: We randomly shuffled the class labels of the 500+500 samples and redo the above experiments. The error rates achieved on the permuted data will demonstrate how well the method can separate the two classes by chance if there is no real classification information in the data. By doing the permutation experiment 500 times, we acquired a rough null distribution of the classification error on shuffled data, from which we estimated the permutation P -value of the error rate obtained from the true data. The left branch of Figure 3 illustrates the procedure of pattern recognition experiment.

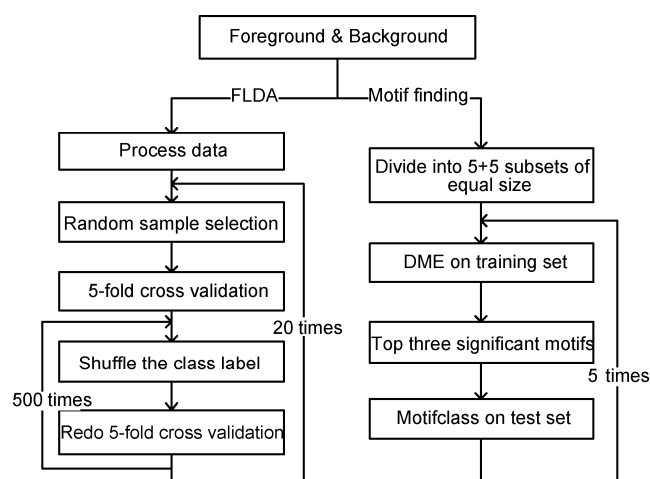


Figure 3 The procedures of FLDA and Motif finding. Left branch (FLDA): (1) process the data to transform each sequence into a normalized feature vector; (2) randomly select 500 samples from foreground and background group, respectively; (3) 5-fold cross validation; (4) permutation test for 500 times; (5) repeat steps 2, 3, 4 for 20 times. Right branch (Motif finding): (1) divide the foreground and background sets into five subsets of equal size, respectively; (2) use DME to identify the top 3 motifs on train set (four subsets); (3) get the top 3 significant motifs; (4) call Motifclass to evaluate the quality of these given motifs on test set (1 subset); (5) repeat steps 2, 3, 4 for five times. We did not do random sampling in motif finding, as foreground and background datasets are not needed to be balanced for DME and Motifclass.

2.2 Motif finding

Sequence motifs are short patterns in a set of DNA sequences that occur more frequently than expected or that appear more frequently in one group of sequences (the foreground) than in another group (the background). They typically correspond to informative signals on the genome such as transcription factor binding sites. We applied motif finding approaches on the sequences of genes with partner antisense transcripts to investigate whether there are special sequence signals enriched in this group compared with the background genes without antisense partners. The right branch of Figure 3 shows the procedures of motif finding analysis.

2.2.1 DME and Motifclass

DME (discriminating matrix enumerator) is a widely used tool for identifying motifs overrepresented in the set of foreground sequences relative to a set of background sequences [12]. It uses an enumerative algorithm to exhaustively and efficiently search a discrete space of matrices, scoring each matrix according to its relative overrepresentation in the foreground. After refinement, the highest scoring matrices will be reported as the overrepresented motifs. In our experiments, we tried to find motifs of lengths of 9, 12 and 15, and set the program to output the top 3 motifs in each experiment.

Motifclass is another tool from *CREAD* package [13] for qualifying the efficiency of given motifs in discriminating

two groups of sequences and evaluating the significance of the motifs. It scores each sequence from both foreground and background set. Using the scores, Motifclass reclassifies the foreground and background sequences and compares them with the original classification. The better classification accuracy one acquires, the more significant the given motif is. One can use relative error rate and *P*-value to assess the result. The relative error rate is defined as

$$\text{Relative error rate} = 1 - \frac{\text{sensitivity} + \text{specificity}}{2},$$

where sensitivity is the proportion of true foreground sequences that are classified correctly and specificity is the proportion of true background sequences being classified correctly. Relative error rate is appropriate for the situation of unbalanced foreground and background sets. The *P*-value is obtained by permuting sequence labels 1000 times and calculating the percentage of cases where the error rate on the permuted data is equal to or less than the relative error rate obtained from the true data. We first used DME to identify a list of top-ranking motifs on training set and then used Motifclass to evaluate the significance of those motifs on test set.

2.2.2 Experiment setting

As shown in Figure 4, we designed seven groups of experiments to identify possible sequence signatures, corresponding to different ways of organizing the data. Considering that the motif appearance might be correlated with its position around coding genes, we set three experiments (groups 1, 2 and 5) where the foreground set consists of sequences from different regions around their partner genes. The regulation effect is influenced by the TSS overlapping (see Introduction), therefore we set four experiments (groups 3, 4, 6 and 7) on the four sets, according to the TSS overlapping feature we obtained during pre-processing of the original dataset. Table S1 in Supporting Information further explains the construction rule of each group.

In each experiment, we divided both the foreground and background sets into five equal-sized subsets. Each time one subset was taken as the test set and the other four as the training set.

3 Results

We used *k*-mer sequence features with *k*=2, 3, 4 and 5 in our pattern recognition experiment. Both the training error and test error in the cross-validation experiments were recorded, and compared with test errors in the permutation experiments. When using *k*=2, the features are the relative frequency of all 16 dinucleotides in the sequences. The average training error rate is 41.7% and the average test error rate is 43.5%. There are 29 cases among the 500 permutation experiments that result in error rates equal to or less



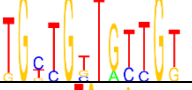
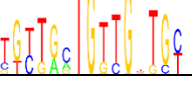


Figure 4 Locations of sequence segments in each foreground and background set for the motif analysis. In all seven groups, foreground sets consist of sequences around primary-antisense pair while background sets are composed of sequences around primary gene without antisense partner. The sequence definition in each group is referred to Figure 3. Notations: the yellow segment refers to the sequences we used in each experiment. Symbol T in groups 3 and 4, and inverted T in groups 6 and 7 represent the TSS region. The number on the top-left corner of each panel is the sequence number in each foreground or background set (see more detailed explanations in the Supporting Information).

than 43.5%, with a P -value of 0.058. When using $k=3$, the average training and test error rates are 34.4% and 40.3%, respectively, with a permutation P -value of 0.004. When $k=4$, the average training and test error rates are 24.5% and 38.7%, respectively, and the permutation P -value is 0.003. When $k=5$, the error rates become larger (40.0% for training and 46.6% for testing), and so is the permutation P -value (0.224). The classification accuracy is not perfect, but the

cross-validation test error rates are statistically significant when $k=3$ and 4. This indicates that the two classes of sequences (genes with and without partner antisense transcripts) are partially separable, and nucleotide compositions contain information to discriminate the two classes. Interestingly, significant classification performance is reached only with $k=3$ and 4. Further analysis is needed to define the possible biological reasons behind this phenomenon.

Table 1 Typical motifs identified in the genes with partner antisense transcripts in *S. cerevisiae*

Motif logo	Relative error rate	Sensitivity	Specificity	P-value
	0.40	0.66	0.54	0.009
	0.39	0.41	0.81	0.002
	0.38	0.75	0.48	< 0.001
	0.37	0.62	0.63	< 0.001

For the motif analysis, most of the groups did not show any significant motif except group 1 (Table 1). The full results of all the seven motif-finding experiments are provided in the Supplementary File. In general, the discriminative power of the identified motifs is not very striking (relative error rate ≈ 0.4). However, a significant motif with consensus sequence 5'-TGTTG-3' was repeatedly identified in different training and test sets of group 1. This motif seems to be evenly distributed along the primary coding genes by showing no significant position preference relative to TSS or gene body. Interestingly, its consensus sequence is similar to the binding site of human poly(ADP-ribose) polymerase (PARP-1) [14]. PARP-1 is an important nuclear enzyme that is conserved in most eukaryotic organisms. Although PARP-1 homolog protein has not been found in yeast [15], it is reported that expressing human PARP1 in *S. cerevisiae* significantly influences the cell status and inhibits cell growth [16], which suggests that this 5'-TGTTG-3' sequence motif might be functional in yeast and be the target of some unknown regulators. Our analysis suggests that this unknown regulatory mechanism might contribute to the antisense transcriptional regulation in yeast.

4 Discussion

The wide existence of antisense transcripts and their regulatory roles on the corresponding protein-coding genes have attracted many researchers' attention. However, sequence signatures of or around the protein-coding genes that may be decisive or informative for the existence of an antisense transcript have not been reported. In this work, we collected all experimentally verified antisense transcripts in the yeast genome to investigate this question. We applied the Fisher's linear discriminant analysis on the *k*-mer sequence features for classifying genes with and without partner antisense transcripts, and further conducted motif analysis for sequences from genes with partner antisense transcripts. The results showed that although the classification of the two

classes with *k*-mer sequence features or with detected motifs are not perfect, the separability of the two classes and the enrichment of some motifs are statistically significant. This indicates that there are some sequence composition patterns associated with the existence of antisense transcripts. Though the sequence patterns identified here could not fully explain the differences of the two groups, the observations opened a new door for the investigation and understanding of antisense transcripts.

In our study, the regions for extracting the sequences for pattern recognition or motif analysis were set arbitrarily and not fine-tuned, and we did not try more comprehensive methods for the pattern recognition and motif analysis. These could be the reasons why the classification accuracy was not very good. Another more important reason for the imperfect classification is that there may be different mechanisms for the existence of antisense transcripts, and different genes may have different types of sequence features or patterns. The high error rates with small *P*-values indicate that this might be the situation. We may need to group all genes into smaller subsets where genes are more homogeneous with each other before performing pattern recognition and motif analysis on the subsets. The functional implications of the identified motif also need to be further studied.

This work was partially supported by the National Basic Research Program of China (2012CB316504 and 2012CB316503) and the National Natural Science Foundation of China (91010016).

- Chen J, Sun M, Kent WJ, Huang X. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res*, 2004, 32: 4812–4820
- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagioli M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminicki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale

- RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 2005, 309: 1559–1563
- 3 Sun M, Hurst LD, Carmichael GG, Chen J. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity. *Genome Res*, 2006, 16: 922–933
 - 4 Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfield EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, de Grey AD, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol*, 2002, 3: RESEARCH0083
 - 5 Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y. Antisense transcripts with rice full-length cDNAs. *Genome Biol*, 2003, 5: R5
 - 6 Jen CH, Michalopoulos I, Westhead DR, Meyer P. Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol*, 2005, 6: R51
 - 7 Wang XJ, Gaasterland T, Chua NH. Genome-wide prediction and identification of *cis*-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biol*, 2005, 6: R30
 - 8 David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA*, 2006, 103: 5320–5325
 - 9 Granovskaia MV, Jensen LJ, Ritchie ME, Toedling J, Ning Y, Bork P, Huber W, Steinmetz LM. Genome Biol High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biology*, 2010, 11: R24
 - 10 Xu Z, Wei W, Gagneur J, Clauder-Münster S, Smolik M, Huber W, Steinmetz LM. Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol*, 2011, 7: 468.1–468.10
 - 11 Zhang X. Pattern Recognition. (3rd ed.) (in Chinese). Beijing: Tsinghua University Press, 2010
 - 12 Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci. USA*, 2005, 102: 1560–1565
 - 13 Smith AD, Sumazin P, Xuan Z, and Zhang MQ. DNA motifs in human and mouse proximal promoters predict tissue specific expression. *Proc Natl Acad Sci. USA*, 2006, 103: 6275–6280
 - 14 Huang K, Tidyman WE, Le KU, Kirsten E, Kun E, Ordahl CP. Analysis of nucleotide sequence-dependent DNA binding of poly(ADP-ribose) polymerase in a purified system. *Biochemistry*, 2004, 43: 217–223
 - 15 Rajawat J, Mir H, Begum R. Differential role of poly(ADP-ribose) polymerase in *D. discoideum* growth and development. *BMC Dev Biol*, 2011, 11: 14
 - 16 Tao Zhihua, Gao Peng, Liu Hung-wen. Studies of the Expression of Human Poly(ADP-ribose) Polymerase-1 in *Saccharomyces cerevisiae* and Identification of PARP-1 Substrates by Yeast Proteome Microarray Screening. *Biochemistry* 2009, 48: 45–11754.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Supporting Information

Part 1 Sequence numbers in each of the seven groups (Table S1)

Part 2 Results of the seven experiments

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.